METRICC: Harnessing Comparable Corpora for Multilingual Lexicon Development

Araceli Alonso, Helena Blancafort, Clément de Groc, Chrystel Million, Geoffrey Williams

METRICC: Harnessing Comparable Corpora for Multilingual Lexicon Development[1]

Araceli Alonso, Helena Blancafort, Clément de Groc, Chrystel Million & Geoffrey Williams

Keywords: *comparable corpora, focused web crawler, collocational networks, multilingual dictionaries, Cultural Heritage lexicon*.

Abstract

Research on comparable corpora has grown in recent years bringing about the possibility of developing multilingual lexicons through the exploitation of comparable corpora to create corpus-driven multilingual dictionaries. To date, this issue has not been widely addressed. This paper focuses on the use of the mechanism of collocational networks proposed by Williams (1998) for exploiting comparable corpora. The paper first provides a description of the METRICC project, which is aimed at the automatically creation of comparable corpora and describes one of the crawlers developed for comparable corpora building, and then discusses the power of collocational networks for multilingual corpus-driven dictionary development.

1. Introduction

In recent years there has been much interest in comparable corpora, especially for creating specialist corpora for translation, terminology and contrastive studies. To date, most studies have focused on the process for building-up comparable corpora, mainly by using crawling techniques, rather than reporting on the different uses and exploitation of this kind of corpora. Several studies have already addressed the issue of comparability of corpora from a statistical point of view so as to see at which point in the crawling process the corpora are, or not comparable (Laviosa 1997, Li and Gaussier 2010). Much of the discussions has been devoted to the selection of seeds to get better comparable texts (Daille and Delpech 2010), but, recent studies demonstrating the techniques being used for creating comparable corpora have rarely been applied multilingually (Kilgarriff et al. 2011).

Concerning the exploitation of comparable corpora, most work has focused on terminological extraction for the selection of new seeds or creation of simple monolingual or bilingual lists of terms (Gaussier et al. 2004, Nakao et al. 2009, Prochasson and Morin 2009). In this sense, the adequacy of the terminologies extracted is still an aspect open to question and there is scarcely any evidence on the use of comparable corpora for building-up real corpus-driven multilingual dictionaries.

METRICC is an ambitious project in that it aims to build specialised comparable corpora automatically using comparability statistics so as to extract lexical data. Led by NLP researchers from the University of Nantes, the project brings together a multidisciplinary team including specialists in NLP, statistics, and web crawling as well as specialists in corpus-driven lexicography and terminology.

This paper focuses on the use of the mechanism of collocational networks proposed by Williams (1998) for exploiting comparable corpora. The study aims to illustrate how collocational networks can be used to extract relevant lexical units related to a specific domain (Alonso et al. 2011) as well as for the selection of the main headwords to compile real corpus-driven multilingual dictionaries.

2. Comparable corpora and multilingual lexicon development

2.1. *State of the art*

Manual compilation of monolingual, bilingual or multilingual lexica and terminologies is extremely time-consuming and costly. As a consequence, research on building automatically monolingual and bilingual and, to a lesser extent, multilingual lexical resources has remained ongoing since the 90s and despite major advances is far from being totally satisfactory. Early work mostly focused on the creation of parallel corpus (Chen 1993, Kay and Röscheisen 1993, Melamed 1997). However, parallel corpora and groups of parallel texts with their corresponding translations remain relatively scarce, especially for specialised domains and for language pairs that do not include English. This lack of resources has motivated research into comparable corpora (Fung and McKeown 1997, Fung and Yee 1998, Déjean et al. 2002, Robitaille et al. 2006, Morin et al. 2007, Morin and Prochasson 2011). These are seen as concerning texts that belong to a same topic or domain, but are not translations of each other. Déjean et al. (2002) define comparable corpora as '*deux corpus de deux langues L1 et L2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue L1, respectivement L2, dont la traduction se trouve dans le corpus de la langue L2, respectivement L1'*. By 'non négligeable' the authors mean that we cannot trace a line between parallel and non parallel corpora, they rather represent a continuum. This reinforces the idea of comparable corpora as a useful source for creating translation memories, and bilingual or multilingual terminologies.

With the increasing amount of textual data available on the net, more and more researchers have worked on the compilation of corpora from the web, a technique known as *web as corpus* (Kilgarriff and Greffenstette 2003). To obtain domain-specific data, focused crawlers — also named thematic or topic crawlers — have been developed to gather comparable corpora for a specific domain by giving domain-specific seeds (terms) as input (WebBootCat, Wüska, etc). A topical web crawler harvests comparable corpora from domain-specific Web portals or using query-based crawling technologies with several types of conditional analysis. However, as stated in the introduction, most research to date has centered on the methods and techniques to build comparable corpora, but the exploitation of these corpora has been scarcely addressed.

Concerning the development of lexicographical resources, some recent work has been done on compiling dictionaries from monolingual corpora which may be broaden up to other languages (Haghighi et al. 2008). Techniques for developing bilingual lexicons from parallel corpora have been also studied (Gale and Church 1991, Fung 1995) as well as different methods to extract lexicons from translation memories (Neff and McCord 1990) or from the web (Nazar et al. 2008). However, studies on compilation of real multilingual dictionaries from comparable corpora have hardly been developed (Bourigault et al. 2001, Teubert 2007), and in most cases research has been addressed to the automatic compilation of lists of words and development of automatic extractor of terms without taking into account the potential of a corpus as a source of information to give account of the use of lexical items.

2.2. *The Metricc Project*

The aim of the French nationally funded METRICC (Translation Memories, Information Retrieval and Comparable Corpora) project is to exploit the possibilities offered by comparable corpora in three specific industrial applications: translation memories, cross-lingual information retrieval and multilingual categorisation. The Project is built around four

main tasks: constructions of comparable corpora, lexicon extraction, application to translation memory, application to cross-lingual information retrieval and multilingual categorisation.

The three year METRICC project, led by the University of Nantes, is financed by the French National Research Agency. Three public laboratories, Lina (*Laboratoire d'Informatique de Nantes Atlantique*), the LIG (*Laboratoire d'Informatique de Grenoble*) and the VALORIA (*Laboratoire de Recherche en Informatique et ses Applications de Vannes et Lorient*), as well as three industrial partners, Lingua et Machina, Sinequa and SYLLABS are participating in the project. At this stage, the main working languages are English and French, though some work is also being developed in other languages such as Japanese or Spanish. More information on the project is available at the website.[2]

The Metricc project is work-in-progress. Most of the research to date has addressed the compilation of comparable corpora. To do this, different crawlers using different techniques have been developed. We are currently assessing the output from these tools, so as to compare the crawlers and the comparable corpora created. Different techniques for improving corpus comparability have also been developed. In relation to lexicon extraction, most of the research has been devoted to the extraction of bilingual lexicons. In this sense, this paper extends the research to the possibility of compiling not only bilingual lexicons, but multilingual ones.

3. Collocational networks for compiling multilingual organic dictionaries

In this study is hypothesised that the mechanism of collocational networks (Williams 1998) may be a potential tool for exploiting the comparable corpus and compile a multilingual lexicon related to Cultural Heritage. The idea of collocational networks is not new and has been put forward and revised for the creation of the *E-Advanced Learner's Dictionary of English Verbs in Science DicSci* (Williams 2006, forthcoming, Williams and Millon 2008a, Alonso et al. 2011). The methodology has also been adopted in other projects (Magnusson and Vanharanta 2003, Järvi et al. 2004, Alonso forthcoming).

Collocational networks are the core element of a methodology, both theoretical and practical in nature, proposed by Williams (1998, 2002) for corpus-driven dictionary building. The mechanism of collocational networks is complemented by that of *collocational resonance* (Williams 2008b, Williams and Millon 2009) and the *Corpus Pattern Analysis* technique developed by Hanks inside his *Theory of Norms and Exploitations* (Hanks 2004, 2006, forthcoming). From a theoretical perspective, collocational networks have been influenced by Sinclair's insights into collocations and the idiom principle (Sinclair 1991), the theory of Lexical Priming proposed by Hoey (2005) and the work on pattern grammar by Hunston and Francis (1999). It also considers Wittgenstein's approach to prototypes (1953), the study on semantic prosody by Louw (1993, 2000|2008), the work on scientific texts by Roe (1977) and the later studies of phraseological aspects of scientific texts developed by Gledhill (2000). A detailed description of the methodology is shown in Williams (1998) and Alonso et al. (2011).

Collocational networks are defined as statistically based chains of collocations, a web of interlocking conceptual clusters realised in the form of words linked through the process of collocation. The idea that collocations "cluster" forming interwoven meaning networks comes from Phillips (1985). Phillips's aim was the study of metastructure within texts and the notion of 'aboutness'. Following this lead, Williams hypothesised that 'the patterns of co-occurrence forming the collocational networks will be unique to any one sublanguage and serve to define the frames of reference within that sublanguage' (Williams 1998: 157). In previous works (Williams 1998, Williams and Millon 2010), it has also been stated that collocational networks not only demonstrate thematic patterns, but they also show the most significant lexical units which out of the analysis of monolingual corpora form the main cognitive nodes

of a specific corpus. The studies developed show that these chains of collocations constitute a powerful tool for headword selection.

Collocational networks method grew and was applied for compiling monolingual dictionaries before being adapted more recently to a multilingual environment, principally through a procedure developed during the IntUne project.[3] In both cases, the advantage of networks arises from an analysis of the lexical environment of words rather than just their discrete usage or even remaining within the constraints of a Keyword in Context span of variable width. Collocational networks enable the analyst to look at the immediate environment of a search word, but then link outwards to the wider meaning context enabling the isolation of lexical units in the Sinclairian sense (Williams 2010).

Bilingual dictionaries, particularly those used in NLP applications, tend to be based on lists of equivalents, or near equivalents found by translating and, and possibly verifying in a corpus, from L1 to L2. The lexicons are thus pre-established and the methodology essentially corpus-based. Collocational networks on the other hand are corpus-driven (Tognini-Bonelli 2001). They explore the lexical environment bringing in new words for new contexts, link to a mulilingual crawler, they thus provide a powerful means of building a multilingual lexicon. As has been shown elsewhere (Williams 2002), networks can be used to categorise, and thus organise data conceptually for dictionary building (Alonso et al. 2011, Williams forthcoming). Multilingual networking essentially entrails 'crawling' the two or more corpora from common agreed seed words, the results is this a growing lexicon with comparable categorisations linked to a natural language based ontology. This means that we can not only find equivalents but also see what they mean in context. This contextual meaning is vital as simple surface equivalence can hide important connotive differences between languages that can only be safely linked through lexicographical prototypes (Hanks 1994, 2000) being adapted to a multilingual usage (Williams 2010, Williams et al. 2012).

4. Compiling a multilingual lexicon on Cultural Heritage

4.1. *Cultural Heritage as an example of domain specificity*

In this study, the comparable corpus created is related to Cultural Heritage, as it is one of the domains considered. The interest on Cultural Heritage derives from previous research developed by one of the research groups involved.[4]

'Cultural heritage' is a concept which has changed through time. At one time, it referred exclusively to the monumental remains of cultures, that is, more in the sense of 'built heritage'. 'Cultural heritage' as a concept has gradually broadened its scope to include new categories such as the intangible, ethnographic or industrial heritage. As defined by UNESCO, 'cultural heritage' is an open concept reflecting living culture as much as that of the past. Taking the definition given by the *Donald Horne Institute for Cultural Heritage*[5], 'cultural heritage' can be defined as 'the things, places and practices that define who we are as individuals, as communities, as nations or civilisations and as a species. It is that which we want to keep, share and pass on'. As a field, Cultural Heritage has emerged over the past years and can be seen as an open interdisciplinary domain related to conventional disciplines such as history, anthropology, archaeology, architecture, art history, theology, literature, linguistics, among others.

Thus, the complexity for obtaining texts related to this domain and deciding on what it is a specific term related to Cultural Heritage or not to use as seed is greater than in other more easily defined fields as Medicine or Chemistry. Despite the fuzzy boundaries of the domain being a potential disadvantage, they are in reality an advantage in testing the capacity of the

techniques developed to create comparable corpora as it involves seeing disciplines as interdisciplinary objects rather than closed well-defined fields of knowledge.

4.2. *Creating the comparable corpus on Cultural Heritage by using Babouk*

For the compilation of the corpus on Cultural Heritage, we used the focused crawler Babouk (de Groc 2011), developed by Syllabs in the context of the European financed project TTC (Translation, Terminology and Comparable Corpora).[6] Babouk is a focused web crawler (Chakrabarti et al. 1999) to gather specialised corpora from the web. Babouk's goal is to gather as many relevant webpages as possible on a specialised domain defined by the user by means of seeds.

When crawling with Babouk, a user typically defines a *crawl job* which is a crawling process configured to the user's needs. The crawling process can either start from a set of specific seed terms (domain-specific 'keywords') or seed URLs. Seed terms are usually terms representative of the domain for which the web documents are retrieved. The seed terms are in fact transformed into seed URLs: they are first combined as tuples and submitted as queries to a search engine. The resulting top-ranked URLs are then selected as seed URLs. Once the seed URLs have been chosen or bootstrapped, the crawling process starts. The crawler downloads the first webpage in queue and analyses its relevance given the crawl job's topic. If the webpage is found to be relevant, all of its links are extracted and added to the crawl queue. Otherwise, the webpage is discarded. This process is iterated until a stopping criterion is met or no more relevant documents are found.

The relevance analysis is achieved using a *thematic filter*. The *thematic filter* is composed of a weighted-lexicon-based categoriser built automatically during the first iteration of the crawling process: first, the seeds defined by the user are expanded to a large lexicon using the BootCaT procedure (Baroni and Bernardini 2004). The resulting lexicon is then weighted automatically using a novel *representativity* measure (de Groc et al. 2012). The tool includes the option to visualise and/or download the lexicon. This thematic filter is then used by the categoriser of the crawler. The categoriser allows the crawler to categorise the documents found on the web and uses the thematic filter to compute the relevance of webpages and filter out non relevant documents. Compared to existing focused web crawlers that rely either on machine learning techniques (Chakrabarti et al. 1999) or manually crafted lexicons (Pecina et al. 2011), we believe our approach is an interesting tradeoff that avoids the burden of defining thematic filters manually while providing users with control and understanding of the categorisation process.

While general web crawlers rely on a simple breadth-first search strategy, focused crawlers prioritise their fetch queue in order to download most relevant webpages first (a process called "crawl frontier ordering" in the crawling literature (Cho et al. 1998)). Babouk uses the relevance score of the webpages as given by its categoriser to rank its URLs queue in a way similar to the OPIC criterion (Abiteboul et al. 2003).

Crawling the web is a recursive process that will solely stop when no more relevant documents are found. While this strategy is theoretically sound, the crawl duration might still be very long. This is why Babouk includes several stopping criteria: users can specify the minimum and/or maximum size of the document to be retrieved (number of words or HTML kylobytes size), a maximum crawl depth or even an upper bound time limit. Moreover, a live content-based web-spam filter is applied. Finally, users can limit the crawl to specific domains or file formats (such as Microsoft Office, Open Office, Adobe PDF, or HTML) and apply a blacklist of unwanted URLs or Internet domains.

Once the crawling process is done, Babouk delivers the set of crawled documents in their original format (html, doc or pdf documents) and two additional files for each retrieve file:

- A Dublin Core[7] metadata file characterising each crawled document retained for the corpus with metadata about the crawled documents including their title, original URL, fetch time and language.
- A text file, containing the plain text extracted from the corresponding web page. If the document was originally an HTML webpage, then all boilerplates and HTML mark-ups are removed using the BodyTextExtraction algorithm (Finn et al. 2011).

One of the unresolved key questions when building comparable corpora is the selection of the corresponding seeds to demarcate a fuzzy domain such as that of Cultural Heritage. Two main options are usually considered: manual selection and lexicographical selection. This is, whether the selection is done manually by the user or by looking-up a dictionary. Both options bring about problems, as is noted by Kilgarriff et al. (2011: 123-124).

In order to select the best seeds to get the most adequate comparable corpora, three different crawls per language were launched with different lists of seeds. In this case, our study is based only on English and French, as the methodology and procedure would be the same in case of more languages:

- The first list consisted of domain-specific terms selected manually by a linguist from a corpus gathered manually from the web about Cultural Heritage. This task was performed separately for each language. – e.g. Sample of the initial seed list in English: *built heritage*, *environment heritage*, *national heritage*, *expenditure on heritage*.
- The second list is a parallel list with a selection of the seeds from the first list and their equivalents. – e.g. Sample of the parallel seed list: *world heritage-patrimoine mondial*, *natural heritage-patrimoine naturel*, *industrial heritage-patrimoine industriel*, *heritage conservation-consérvation du patrimoine*.
- The third list was generated automatically from the corpora. For the automatic extraction, an in-house rule-based tool for information extraction was used to extract all simple and complex nouns from the text (de Groc 2011). Results were ranked by frequency and the top-ten resulting nouns were taken into consideration. The procedure was the same for both languages. – e.g. Sample of the weighted seed list: *ancient monuments*, *archaeological sites*, *conservation areas*, *English heritage*, *historic buildings*.

Results obtained from each crawl job revealed that using more seeds did not mean better results. In fact, the crawl job based on the parallel seed list, which contains less seeds than the other two lists, obtained more relevant texts related to Cultural Heritage. As Cultural Heritage is a fuzzy domain, it is necessary to evaluate the different comparable corpora obtained and choose one for the exploitation process. In order to estimate the domain specificity of the three comparable corpora obtained by each of the crawl jobs, a test consisting on an evaluation of the corpus coverage in relation to a reference term list of the domain was run.

For the test, we used the automatic terminological extractor of Syllabs. This pattern-based terminological extractor first selects simple and multi-word term candidates and then ranks them using the relative frequency of a term as suggested by Ahmad et al. (1992). The relative frequency of a term is computed using its frequency of occurrence in the specific as well as in the generic corpus. In this case a general corpus of fifteen million tokens was considered. For the evaluation, we compared the term candidate list obtained automatically for each crawled

corpus to a reference term list in the domain of Cultural Heritage. This reference term list includes 4451 terms and was compiled manually using different resources from the Internet: *The Heritage Conservation Glossary*[8], *Le Répertoire canadien des lieux patrimoniaux*[9], *Le glossaire vocabulaire du patrimoine*[10] and *Le glossaire du patrimoine culturel immatériel de l'Unesco*.[11] The evaluation script calculates the number of exact matches per lemma and form as well as the approximate matches. The approximate matches are calculated using the Levenshtein distance (Nazarenko and Zargayouna 2009). The results obtained for both languages are similar.

Table 1. Results of the corpus coverage for French depending on the seed terms used.

	Manual Seeds	Parallel Seeds	Weighted Seeds
Term list	4451	4451	4451
Output list	39865	143204	118966
Perfect lemma match	932	1167	1123
Perfect form match	637	881	822
Approximate lemma match	269	352	330
Approximate form match	248	320	310
Perfect lemma and form match	1569	**2048**	1945
Perfect and approx match	2086	**2720**	2585
No match	2609	2047	2172

For both languages, the crawl job run using the parallel seeds gave a higher recall than crawls using the other seeds. As illustrated in table 1, a higher number of perfect and approximate matches was achieved. As a result, the comparable corpus chosen for the study presented here is the comparable corpus that was crawled using the parallel seeds as input. The resulting corpus in English has a total of 3,071,041 tokens, while the French one contains 3,625,978 tokens.

4.3. *Compiling the multilingual lexicon on Cultural Heritage*

Once the comparable corpora on Cultural Heritage had been created and tested, collocational networks for each of the languages were generated. Even though our main working languages are English, French and Spanish, in this study only English and French are considered. The methodology and procedure explained would be the same in case when adding more languages.

The two corpora were launched in the Word Sketch Engine tool[12] in order to build the collocational networks using specific grammatical relations. In this case, the study is based in one of the most nuclear lexical units related to Cultural Heritage, *heritage* with 24,558 occurrences in the Babouk_Enparallelseeds Corpus and its correspondent equivalent in French, *patrimoine*, with 21,568 occurrences in the corresponding French corpus.

In the collocational networks shown, the VERB + NOUN_object pattern is involved. For the first level of collocations only the ten most significant verbs according to the salience measure in the pattern VERB + *heritage* (or VERB + *patrimoine*) are taken into account (red nodes in the networks). The second level concerns the ten most significant nouns according to the salience measure for each verb of the first level; these nominal nodes are either in green in the network if they are shared by at least two verbs — for example, within the French collocational network, the nominal node *valeur* is shared by four verbs, namely: *sauvegarder*, *préserver*, *protéger*, and *menacer* —, or in grey if not. In Figures 1 and 2, the collocational networks of, respectively, *heritage* and *patrimoine* extracted from the second crawl by using manual parallel seeds are illustrated as an example.

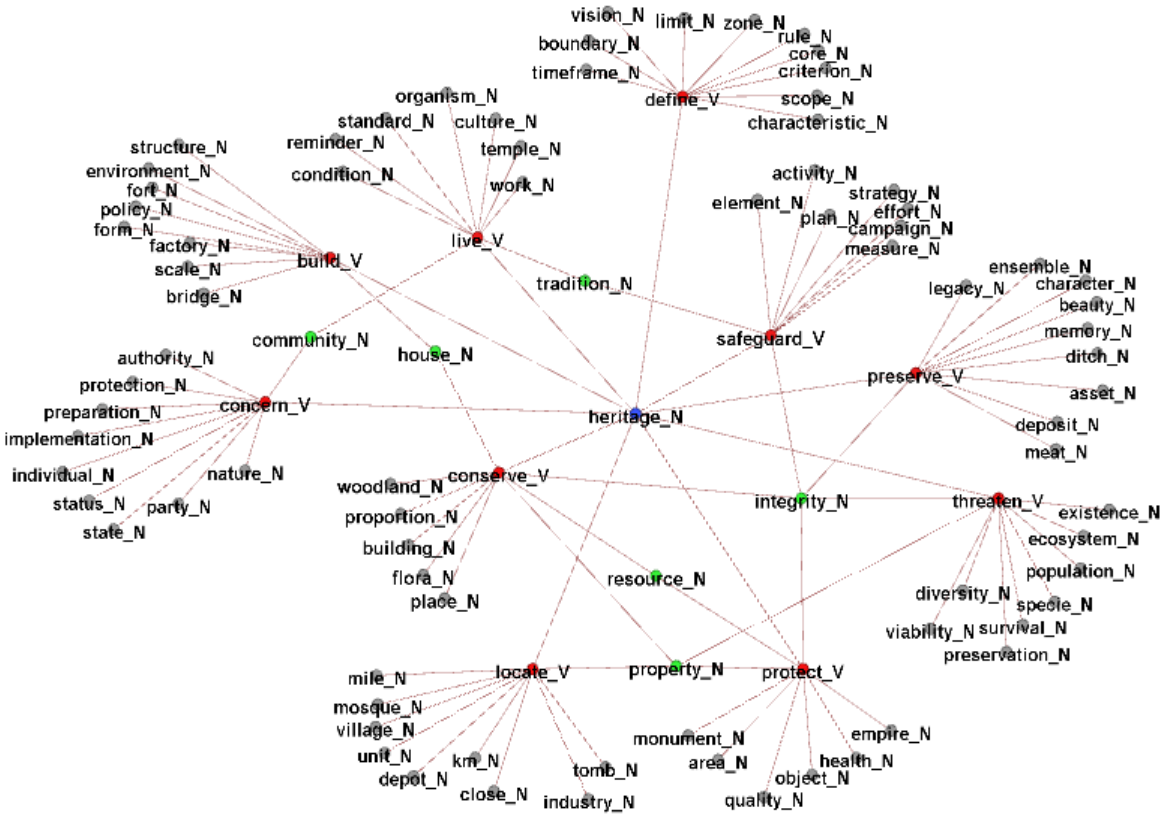

Figure 1. First two-level of collocational network of *heritage* from the BaboukEN_parallelseeds Corpus.

By looking at the collocational network of *heritage* in the BaboukEN_parallelseeds Corpus the most salient verbal collocates are shown – e.g *build* (230 occurrences), *protect* (108), *safeguard* (44), *preserve* (36), *conserve* (28), *define* (29), *concern* (26), *threaten* (15), *locate* (15) and *live* (14).[13] At the same time, the second level of the collocational network shows other significant nouns which collocate with the verbal collocates – e.g. *area* (93), *integrity* (12), *resource* (19), which are collocates of the verb *protect*. Furthermore, the collocational network also highlights those nouns which are shared by more than one verb – e.g. *resource* is a collocate both of *protect* and *conserve*, and *property* is shared by *conserve*, *locate*, *protect* and *threaten*. Going even further, collocates also stresses the differences between verbs which may be considered synonyms or partial synonyms. For instance, what it is conserved is a *building* but also a *place*, *flora* or *woodland*, while what is preserved is the *character*, the *beauty*, the *memory*, the *legacy*. *Conserve* seems to be applied to 'Physical objects', 'Locations' or 'Resources' while *preserve* is applied to more 'Abstract' or 'Intangible things'.

The collocational network highlights not only the most frequent lexical units used in Cultural Heritage, but the most salient ones; those lexical units which form cognitive nodes and which bring meaning to the text. It also shows the relations between the units. In addition, the lexicographical analysis allows the grouping of lexical units into conceptual classes.[14] In this case, *heritage* is mainly used with verbs which fulfil the task of PROTECTING – e.g. *protect*, *safeward* – or the task of LOCATING –e.g. *locate*. Conceptual classes help to clarify usage to the user and also to organise the lexicon not only alphabetically, but also onomasiologically.

At this stage, if we need to create not just a lexicon but a complete corpus-driven dictionary with definitions, examples, etc. the methodology proposed for the DicSci and the above mentioned procedure would be applied. That is, the concordances for each of the

collocates would be analysed in detail in order to create the different patterns of use by applying *Corpus Pattern Analysis* technique.

The same procedure could be applied to the collocational networks based on the other grammatical relations, such as, for instance, the ADJ + NOUN_modifier relation in order to extract the most significant adjectives which are used with *heritage*.

Once the collocational networks of the English corpus have been analysed and the lexical units selected, the same procedure is applied to the French language corpus, and the other possible languages which may be used to extract the multilingual lexicon. Figure 2 shows the corresponding collocational network related to the grammatical relation VERB + NOUN_object in French.

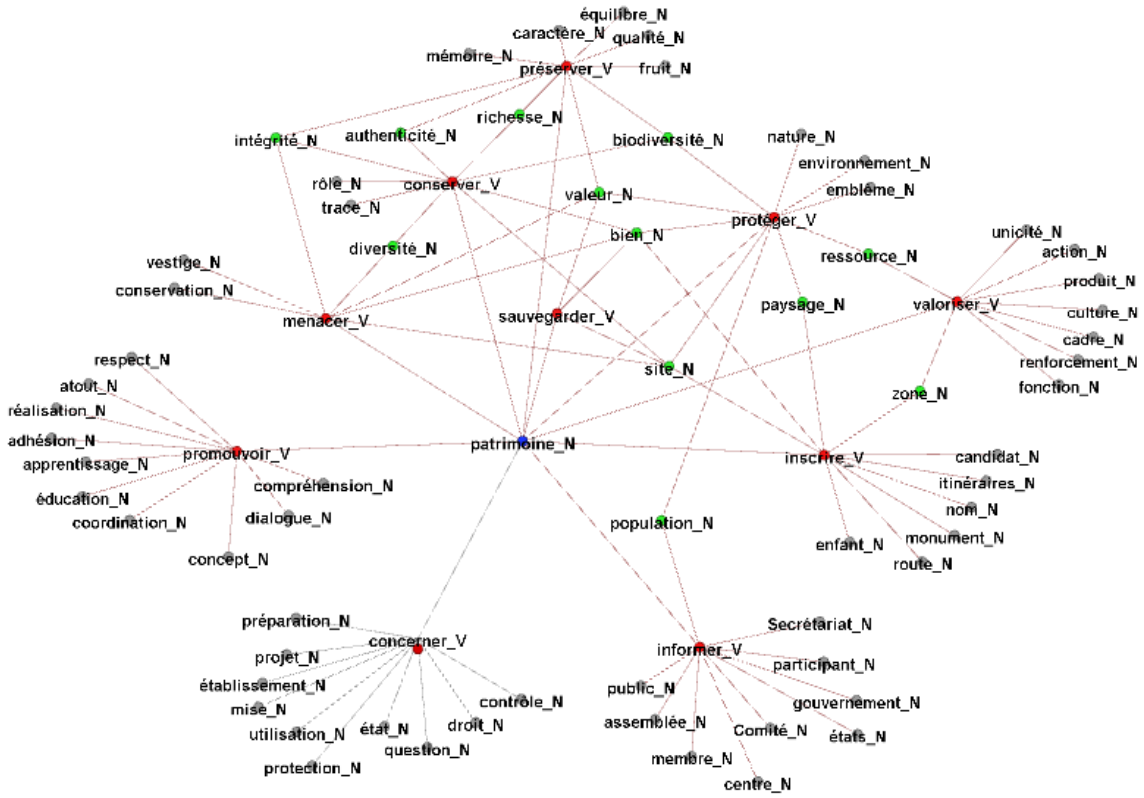

Figure 2. First two-level of the collocational network of *patrimoine* from the BaboukFR_parallelseeds Corpus.

The ten most significant verbal collocates of *patrimoine* ordered by salience are *concerner* (86 occurrences), *protéger* (38), *préserver* (33), *valoriser* (28), *conserver* (18), *menacer* (14), *informer* (13), *sauvegarder* (12), *inscrire* (23), *promouvoir* (11). As in English, there are some collocates which are shared by different verbs – e.g. *richesse*, *authenticité* and *intégrité* are collocates of *préserver* and *conserver*, and *intégrité* is also a collocate of *menacer*. It can be observed that not all verbs are coincident to those used in English, even though there are some coincidences – e.g. *protéger*, *préserver*, *conserver*, *menacer*, *sauvegarder*.

In relation to the collocates, some similarities and differences between the two languages can be noticed. For instance, *préserver* collocates with *intégrité* (10), *qualité* (13), *biodiversité* (10), *fruit* (3), *équilibre* (4), *caractère* (6), *richesse* (4), *authenticité* (3), *mémoire* (3) and *valeur* (14). As in English most of the collocates refer to 'Abstract' or 'Intangible things', although some collocates related to 'Resources' are displayed – e.g. *biodiversity*. It must be stated that the use of *fruit* as one of the most salient collocates of *préserver* in the context of Cultural Heritage is at first glance strange. Looking at the concordances evidences that, in this

case, it is not used as in the sense of 'Food' but as a metaphorical use in the context of '*préserver les fruits de la civilisation de l'homme*'. The use of the verb *conserver*, in contrast, is more similar to that of *préserver* than is the case of *conserve* in English, as the most significant collocates not only refer to 'Locations' –e.g. *site* – or 'Resources' – e.g. *diversité, biodiversité* – but also to 'Abstract' or 'Intangible objects' – e.g. *intégrité, authenticité*. In relation to the functional groups, the main group is also that of PROTECTING.

As in English, the other grammatical relations should also be analysed in order to have a whole picture of the environment of use of *patrimoine* in Cultural Heritage.

In this paper, we have shown only two levels of the collocational networks, but the networks should be broadened to take in more levels until reach a point where collocates are repeated. Thus, the entire process would be applied again for the rest of the most frequent lexical units. It should be noted that every time a collocational network is added, the information may affect the previous networks bringing out new data. In this sense that collocational networks are said to be applicable to the creation of 'organic' dictionaries, in the sense that they grow naturally from the data.

Once we get to this point, the most significant lexical units for each language can be extracted, so that a parallel list of lexical items is created. However, this would mean to just create a parallel word rather than taking into account the use of the units in context. As can be seen from the analyse, the mechanism of collocational networks is a powerful one and allows the development of lexicons which demonstrate the use of lexical units in specialised texts and a means of comparing these uses between different languages. To take this research further, the collocational networks would be used to supply the headwords for a dictionary where one collocational network is linked to the corresponding collocational network in the other language. In this way, the user would have a real picture of the environment of a word in each of the languages.

5. Concluding remarks

The development of research on specialised comparable corpora has been dedicated to showing how to choose significant seeds to create a corpus that is as comparable as possible. Many studies have focused on domains such as Medicine or the like and generate comparable corpora that take into account manual or automatic lists of terms as input for the crawler. However, new domains bring about new needs. In social and interdisciplinary domains difficulties arise for term recognition, as the terminological status of some lexical units is not always clear as, following Hanks (2010), many relevant lexical units have a more phraseological tendency than a terminological one. In reality, most studies ignore the fact that scientific meaning is created in context and, therefore, the importance lies in determining the most significant lexical units which bring meaning to the domain and not in deciding whether a lexical unit is or is not a term. In order to illustrate the important of the lexical environment, the mechanism of collocational networks has been adopted.

Collocational networks show the most significant cognitive nodes of the corpus created which can be considered as the main entries of a multilingual specialised dictionary on Cultural Heritage. They also show differences that can be found between languages as the conceptualisation of the domain from one language to the other may vary. Finally, the observation of concordances of the collocations and collocates illustrated by the collocational networks shows patterns of usage for each unit and allows the comparison of patterns between languages. This information is also considered for creating dictionary entries and may be of extremely importance in building-up multilingual specialised dictionaries (Alonso et al.

2011). It is here that we find information about the use of lexical units in contexts, this being useful not only for decoding, but also for encoding tasks.

Notes

[1] Research for this article was funded by the Equipe LiCoRN of the HCTI research group from the University of Bretagne Sud, the ANR research project Metricc (ANR-08-CORD-013) and the Spanish Ministry of Education as part of the *National Mobility Programme of Human Resources of the R+D National Programme 2008-2011* which has made possible the post-doctoral work of one of the authors.
[2] htpp://www.metricc.com
[3] An introduction to the IntUne project can be found at the website http://www.intune.it
[4] The EC funded IntUne project and PATH, an FP7 proposal that is currently being reworked.
[5] http://www.canberra.edu.au/centres/donald-horne
[6] An introduction to the TTC project can be found at the website http://www.ttc-project.eu
[7] http://dublincore.org
[8] http://www.icomos.org/~fleblanc/documents/terminology/doc_terminology_glossary_ef.html
[9] http://www.historicplaces.ca/fr/pages/about-apropos.aspx
[10] http://langues.univ-paris1.fr/glossairepatrimoinefrancais-anglais.pdf
[11] http://www.unesco.org/culture/ich/doc/src/00265.pdf
[12] For a detailed description of the Word Sketch tool, see Kilgarriff et al. (2004). The tool is available at http://www.sketchengine.co.uk
[13] The collocates *build* and *live* are not really used as verbs, but adopt an adjectival function in the constructions 'building heritage' and 'living heritage', respectively.
[14] For more information on the grouping function of collocational networks, see Alonso et al. (2011) and Williams (forthcoming).

References

Abiteboul, S., M. Preda and G. Cobena 2003. 'Adaptive On-line Page Importance Computation.' In *WWW '03 Proceedings of the 12th International Conference on World Wide Web*. New York: ACM Press, 280–290.

Alonso, A. Forthcoming. 'La caractérisation du lexique de l'environnement à travers des réseaux collocationnels.' *Le Discours et la langue*.

Alonso, A., C. Millon and G. Williams 2011. 'Collocational Networks and their Application to an E-Advanced Learner's Dictionary of Verbs in Science (DicSci).' In I. Kosem and K. Kosem (eds.), *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex 2011*. Ljubljana: Trojina, 12–22. http://www.trojina.si/elex2011/Vsebine/proceedings/eLex2011-2.

Baroni, M. and S. Bernardini 2004. 'BootCaT: Bootstrapping Corpora and Terms from the Web'. In *Proceedings of LREC 2004.* Lisbon: ELDA, 1313–1316.

Bourigault, D., C. Jacquemin and M. C. L'Homme 2001. *Recent Advances in Computational Terminology*. Amsterdam: Benjamins.

Chakrabarti, S., M. van den Berg and B. Dom 1999. 'Focused Crawling: a New Approach to Topic-Specific Web Resource Discovery.' *Computer Networks* 31.11-16: 1623–1640.

Chen, S. F. 1993. 'Aligning Sentences in Bilingual Corpora Using Lexical Information.' In L. K. Schubert (ed.), *31st Annual Meeting of the Association for Computational Linguistics*, 22-26 June 1993, Ohio State University. Columbus, Ohio, USA: Ohio State University, ACL, 9–16.

Cho, J., H. García-Molina and L. Page 1998. 'Efficient Crawling through URL Ordering.' *Computer Networks and ISDN Systems* 30: 161–172.

Daille, B. and E. Delpech 2010. 'Dealing with Lexicon Acquired from Comparable Corpora: Validation and Exchange.' In *Proceedings 9th Conference on Terminology and Knowledge Engineering (TKE)*. Ireland: Fiontar, Dublin City University.

de Groc, C. 2011. 'Babouk: Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction.' In *IEEE/WIC/ACM 2011: International Conference on Web Intelligence and Intelligent Agent Technology*, August 22-27, 2011, Campus Scientifique de la Doua, Lyon, France.

de Groc, C., X. Tannier and C. de Loupy 2012. 'Un critère de cohésion thématique et son application à la création de terminologies bilingues.' In *Proceedings of the Conférence du Traitement Automatique des Langues Naturelles*.

Déjean, H., E. Gaussier and F. Sadat 2002. 'An Approach based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction.' In *COLING 2002, Proceedings of the 19th International Conference on Computational Linguistics*. Taipei, Taiwan: Howard International House and Academia Sinica, 1–7.

Finn, A., N. Kushmerick and B. Smyth 2011. 'Fact or fiction: Content Classification for Digital Libraries.' In *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries.*

Fung, P. 1995. 'A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora.' In H. Uszkoreit (ed.), *ACL 1995, 33rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 26-30 June 1995. MIT, Cambridge, Massachusetts, USA. Morgan Kaufmann Publishers, ACL, 236–243.

Fung, P. and K. McKeown 1997. 'Finding Terminology Translations from Non-parallel Corpora.' In *Proceedings of the 5th Annual Workshop on Very Large Corpora*. Hong Kong, 192–202.

Fung, P. and Y. L. Yee 1998. 'An IR Approach for Translating New Words from Nonparallel, Comparable Texts'. In C. Boitet and P. Whitelock (eds.), *COLING-ACL'98, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, August 10-14, University of Montreal, Montreal. Morgan Kaufmann Publishers, ACL, 414–420.

Gale, W. A. and K. W. Church 1991. 'A Program for Aligning Sentences in Bilingual Corpora.' In D. E. Appelt (ed.), *29th Annual Meeting of the Association for Computational Linguistics*, 18-21 June 1991, University of California, Berkeley. California, USA: University of California, ACL, 177–184.

Gaussier, E., J. M. Renders, I. Matveeva, C. Goutte and H. Déjean 2004. 'A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora.' In D. Scott, W. Daelemans and M. A. Walker (eds.), *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 21-26 July, Barcelona, Spain, 526–533.

Gledhill, C. (2000). *Collocations in Science Writing*. Tübingen: Gunter Narr Verlag.

Haghighi, A., P. Liang, T. Berg-Kirkpatrick and D. Klein 2008. 'Learning Bilingual Lexicons from Monolingual Corpora.' In K. McKeown, J. D. Moore, S. Teufel, J. Allan and S. Furui (eds.), *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, June 15-20. Columbus, Ohio, USA. ACL, 771–779.

Hanks, P. 1994. 'Linguistic Norms and Pragmatic Exploitations or, Why lexicographers need Prototype Theory, and vice versa.' In F. Kiefer, G. Kiss, and J. Pajzs (eds.), *Papers in Computational Lexicography: Complex 94*. Budapest: Hungarian Academy of Sciences, 89–113.

Hanks, P. 2000. 'Do Word Meanings Exist?' *Computers and the Humanities* 34.1-2: 205–215.

Hanks, P. 2004. 'The Syntagmatics of Metaphor and Idiom.' *International Journal of Lexicography* 17.3: 245–274.

Hanks, P. 2006. 'The Organization of the Lexicon: Semantic Types and Lexical Sets.' In E. Corino, C. Marello and C. Onesti (eds.), *Atti del XII Congresso Internazionale di Lessicografia : Torino, 6-9 settembre 2006*. Alessandria: Edizioni dell'Orso, 1165–1168.

Hanks, P. 2010. 'Terminology, Phraseology, and Lexicography.' In A. Dykstra and T. Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress, Leeuwarden, 6-10 July 2010*. Ljouwert: Fryske Akademy / Afuk.

Hanks, P. Forthcoming. *Lexical Analysis: Norms and Exploitations*. Massachusetts: The MIT Press.

Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hunston, S. and G. Francis 1999. *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam and Philadelphia: John Benjamins.

Järvi, P., C. Magnusson and H. Vanharanta 2004. 'The Evaluation of the Collocational Method in the Marketing Context.' *Proceedings of the 2004 ABAS International Conferences*. http://www.sba.muohio.edu/abas/2004/proceedings.html

Kay, M. and M. Röscheisen 1993. 'Text Translation Alignment.' *Computational Linguistics* 19.1: 121–142.

Kilgarriff, A., P. V. S. Avinesh and J. Pomikálek 2011. 'Comparable Corpora BootCaT.' In I. Kossem and K. Kossem (eds.), *Electronic Lexicography in the* 21st *Century: New Applications for New Users. Proceedings of eLex 2011*. Ljubljana: Trojina, 122–128.

Kilgarriff, A. and G. Grefenstette 2003. 'Introduction to the Special Issue on Web as a Corpus.' *Computational Linguistics* 29.3.

Kilgarriff, A., P. Rychly and D. Tugwell 2004. 'The Sketch Engine.' In G. Williams and S.Vessier (eds.), *Proceedings of the eleventh EURALEX International Congress EURALEX 2004 Lorient, France, July 6-10, 2004*. Lorient: Université de Bretagne-Sud, 105–116.

Laviosa, S. 1997. 'How Comparable can 'comparable' Corpora Be?' *Target* 9.2: 289–319.

Li, B. and E. Gaussier 2010. 'Improving Corpus Comparability for Bilingual Lexicon Extraction from Comparable Corpora.' In Ch. Huang and D. Jurafsky (eds.), *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference,* August 23-27. Beijing, China: Tsinghua University Press, 644–652.

Louw, B. 1993. 'Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies.' In M. Baker (ed.), *Text and Technology.* Amsterdam: John Benjamins, 157–176.

Louw, B. 2000|2008. 'Contextual Prosody Theory: Bringing Semantic Prosodies to Life.' In C. Heffer and H. Sauntson (eds.), *Words in Context: A Tribute to John Sinclair on his Retirement*. CD-ROM: English Language Research Discourse Analysis Monograph No. 18. Reprinted in online journal *Texto* (2008): http://www.revue-texto.net/index.php?id=124

Magnusson, C. and H. Vanharanta 2003. 'Visualizing Sequences of Texts Using Collocational Networks.' In P. Perner and A. Rosenfeld (eds.), *Proceedings MLDM'2003, Proceedings of the 3rd International Conference on Machine Learning and Data Mining in Pattern Recognition*. Heidelberg: Springer-Verlag Berlin. 276–283.

Melamed, I. D. 1997. 'A Portable Algorithm for Mapping Bitext Correspondence.' In P. R. Cohen and W. Wahlster (eds.), *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, July 7-12, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain. Morgan Kaufmann Publishers, ACL, 305–312.

Morin, E., B. Daille, K. Takeuchi and K. Kageura 2007. 'Bilingual Terminologymining - Using Brain, not Brawn Comparable Corpora'. In J. A. Carroll, A. van den Bosch, A. Zaenen (eds.), *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, June 23-30, Prague, Czech Republic. ACL, 664–671.

Morin, E. and E. Prochasson 2011. 'Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora.' In *4th Workshop on Building and Using Comparable Corpora (BUCC 2011)*, June 24, Portland, Oregon, USA.

Nakao, Y., L. Goeuriot and B. Daille 2009. 'Multilingual Modalities for Specialized Languages.' *Terminology* 16.1: 77–106.

Neff, M. and M. McCord 1990. 'Acquiring Lexical Data from Machine-readable Dictionary Resources for Machine Translation.' In *Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*.

Nazar, R., L. Wanner and J. Vivaldi 2008. 'Two-Step Flow in Bilingual Lexicon Extraction from Unrelated Corpora.' In J. Hutchins and W. V. Hahn (eds.), *Proceedings of the 12th Conference of the European Association for Machine Translation*, September 22-23, Hamburg, Germany. Hamburg: HITEC e.V, 140–149.

Pecina, P.; A. Toral, A. Way, V. Papavassiliou, P. Prokopidis and M. Giagkou, 2011. 'Towards Using Web-crawled Data for Domain Adaptation in Statistical Machine Translation.' In M. L. Forcada, H. Depraetere and V. Vandeghinste (eds.), *Proceedings of the 15th Conference of the European Association for Machine Translation*, May 30-31, Leuven, Belgium. Belgium: Katholieke Universiteit Leuven, EAMT, 297–304.

Phillips, M. (1985). *Aspects of Text Structure: An investigation of the lexical Organisation of Text*. Amsterdam, North Holland: Elsevier Science, Ltd.

Prochasson, E. and E. Morin 2009. 'Points d'ancrage pour l'extraction lexicale bilingue à partir de petits corpus comparables spécialisés.' *Traitement automatique des langues (TAL)* 50.1: 283–304.

Robitaille, X., Y. Sasaki, M. Tonoike, S. Sato, T. Utsuro 2006. 'Compiling French-Japanese Terminologies from the Web.' In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, April 3-7, Trento, Italy. EACL, 225–232.

Roe P. 1977. *The Notion of Difficulty in Scientific Text*. PhD Thesis, University of Birmingham.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Teubert, W. (ed.). 2007. *Text Corpora and Multilingual Lexicography*. Amsterdam: John Benjamins.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work. Studies in Corpus Linguistics*, 6. Amsterdam/Philadelphia. John Benjamins Publishing Company.

Williams, G. 1998. 'Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles.' *International Journal of Corpus Linguistics* 3.1: 151–171.

Williams, G. 2002. 'In Search of Representativity in Specialised Corpora: Categorisation through Collocation.' *International Journal of Corpus Linguistics* 7.1: 43–64.

Williams, G. 2003. 'From Meaning to Words and Back: Corpus Linguistics and Specialised Lexicography.' *Asp, la revue du GERAS* 39-40: 91–106. http://asp.revues.org/1320

Williams, G. 2006. 'Advanced ESP and the Learner's Dictionary.' In E. Corino, C. Marello and C. Onesti (eds.), *Atti del XII Congresso Internazionale di Lessicografia : Torino, 6-9 settembre 2006*. Alessandria: Edizioni dell'Orso, 795–801.

Williams, G. 2008a. 'Verbs of Science and the Learner's Dictionary.' In E. Bernal and J. DeCesaris (eds.), *Proceedings of the XIII Euralex International Congress: Barcelona, 15-19 July 2008*. Barcelona: L'Institut Universitari de Lingüistica Aplicada, Universitat

Pompeu Fabra, 797–806.
Williams, G. 2008b. 'The Good Lord and his Works: A Corpus-based Study of Collocational Resonance.' In S. Granger and F. Meunier (eds.), *Phraseology: an interdisciplinary perspective*. Amsterdam: John Benjamins, 159–174.
Williams, G. 2010. 'A Cultivated Audience: Comparable Corpora and Cross-Language Collocation'. In M. Bondi, S. Cacchiani and G. Palumbo (eds.), *Corpus Linguistics and Language Variation, Special Issue of RILA* (Rassegna Italiana di Linguistica Applicata) Bulzoni.
Williams, G. Forthcoming. 'Bringing Data and Dictionary Together: Real Science in Real Dictionaries.' In A. Bolton, S. Thomas and E. Rowley-Jolivet (eds.), *Corpus-Informed Research and Learning in ESP: Issues and Applications*. Amsterdam: John Benjamins, 219–240.
Williams, G. and C. Millon 2009. 'The General and the Specific: Collocational Resonance of Scientific Language.' In M. Mahlberg, V. González-Díaz and C. Smith (eds.), *Proceedings of the Corpus Linguistics Conference CL2009*, July 20-23. Liverpool: University of Liverpool. http://ucrel.lancs.ac.uk/publications/cl2009/
Williams, G. and C. Millon 2010. 'Going Organic: Building an Experimental Bottom-up Dictionary of Verbs in Science.' In A. Dykstra and T. Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress, Leeuwarden, 6-10 July 2010*. Ljouwert: Fryske Akademy / Afuk, 1251–1257.
Williams G., R. Piazza and D. Giuliani 2012. 'Nation and Supernation: A Tale of Three Europes.' In P. Bayley and G. Williams (eds.), *European Identity: What the Media Say*. Oxford: Oxford University Press, 55–83.
Wittgenstein, L. 1953. *Philosophical investigations*. Oxford: Blackwell.